Why models fail? Characterizing dataset differences through the lens of model desiderata

Anonymous ACL submission

Abstract

Machine learning systems' effectiveness depends on their training data, yet dataset collection remains critically under-examined. Using hate speech detection as a case study, we present a systematic evaluation pipeline examining how dataset characteristics influence three key model desiderata: robustness against distribution shift, satisfaction of fairness criteria, and explainability. The aim is to highlight the importance of a holistic evaluation of the models and the datasets. Through analysis of 21 different corpora, we uncover crucial inter-dependencies between these dimensions that are often overlooked when studied in isolation. We report significant cross-corpus generalization failures and quantify pervasive demographic biases, with 85.7% of datasets generating models exhibiting Group Membership Bias scores near random chance. Our experiments demonstrate that post-hoc explanations exhibit substantial volatility to changes in training distributions, independently from the choice of feature attribution method or model architecture. These explanations also produce inconsistent and contradictory responses when evaluated under distribution shift. Our findings reveal critical though underestimated synergies between training distributions and model behavior, demonstrating that without careful examination of training data characteristics, we risk deploying systems that perpetuate the very harm they are designed to address.

1 Introduction

Data, more than computing advances, has sparked the AI breakthrough. A canonical example lies in facial detection systems; the performance barriers were transcended not through the perceived computational progress in deep learning, but through the availability of vast training data that enabled more robust feature learning (Torralba and Efros, 2011). This fundamental dependency on data presents several open challenges: How do we know what is different between datasets in the same domain? The question surrounding data collection and comparison are of paramount importance, arising in scenarios such as dataset augmentation, multi-source data integration, and distribution shift detection (Babbar et al., 2024). Despite this, dataset collection remains the most under-scrutinized component of the machine learning pipeline, with an estimated 92% of machine learning practitioners encountering data cascades, or downstream problems resulting from poor data quality (Sambasivan et al., 2021).

This study examines how training distributions manifest as differences in downstream model behavior under three key desiderata: robustness against distribution shift, fairness, and explainability. These properties directly operationalize key OECD principles for trustworthy AI (OECD, 2024), yet remain siloed in machine learning research. No existing work jointly quantifies a dataset's responsibility for model trustworthiness across multiple dimensions simultaneously, despite their evident interdependence in real-world deployments (Ethayarajh and Jurafsky, 2020; Mitchell et al., 2019). Our work addresses this research fragmentation through a comprehensive evaluation framework that measures how dataset characteristics influence all three desiderata. To the best of our knowledge, this also represents one of the first investigations into how learned representations shape the reliability of post-hoc explainability methods when evaluated under shift and in conditions of model degradation. This redirection is necessary because these properties may exhibit complex interdependencies: distribution shifts in deployment environments can invalidate fairness guarantees established on in-distribution data, while simultaneously compromising the reliability of explanation methods that practitioners depend on for regulatory compliance and debugging.

We selected 21 hate-speech detection corpora for our analysis as they offer an optimal testbed for

examining trustworthiness across multiple model dimensions. These datasets are publicly available and were collected from the different original publications in MetaHate (Piot et al., 2024). Hate speech detection, while crucial for online safety, faces fundamental challenges in supervised learning approaches. These systems exhibit poor cross-corpus generalization despite operating in shared semantic spaces, demonstrate systematic performance disparities across demographic groups, and employ opaque decision boundaries that often resist interpretation (Arango et al., 2019; Davidson et al., 2019). While our work focuses on hate speech detection, the methodology is domain-agnostic and applicable across any NLP task where robustness, fairness, and explainability are critical concerns.

We present the first integrated framework for evaluating natural language datasets across multiple dimensions of model trustworthiness, showcasing it using widely accepted tools and metrics. In this work, we make the following contributions:

1. We provide empirical of pervasive distributional misalignment in hate speech detection datasets through cross-dataset generalization experiments. The experiments quantify significant performance degradation during out-of-domain evaluation, even among datasets with shared objectives and data sources.

2. We quantify the extent of demographic bias in hate speech detection systems, revealing that 85.7% of evaluated datasets produce models with Group Membership Bias scores approximating random guessing (0.5).

3. We demonstrate that faithfulness of post-hoc explanations may be significantly influenced by training data distribution, independent of model architecture and feature attribution methods. We challenge common assumptions about the relationship between model performance and faithfulness of post-hoc explanations; the inherent explainability of simple models compared to more complex ones; and the reliability of post-hoc explainability methods under distribution shift.

2 Background

The landscape of machine learning research has undergone a fundamental shift, with increasing attention paid to data itself as a key driver of model performance. This spans both theoretical work examining how data distributions affect learning and generalization (Adebayo et al., 2018; Arpit et al., 2017; Badjatiya et al., 2017; Jiang et al., 2019; Yang et al., 2022, 2024), and their influence on model fairness and bias (Dwork et al., 2012; Feldman et al., 2015; Hardt et al., 2016; Romei and Ruggieri, 2014; Zliobaite, 2015). In post-hoc explainability research, Ribeiro et al. (2021) remains the only work investigating the role of data in post-hoc explainability. This increased focus on data has catalyzed practical advances in data-centric machine learning methodologies (DMLR, 2024), with multiple research threads emerging around dataset construction (Almohaimeed et al., 2023; Mosquera Gómez et al., 2023; Pingle et al., 2023; Shinde et al., 2024) and the application of these approaches to new domains (Arnaiz-Rodriguez and Oliver, 2024; Deng and Ma, 2024; Kohli et al., 2024; Vysogorets and Kempe, 2024; Zhao et al., 2024). Simultaneously, it has prompted crucial discussions around ethical frameworks governing AI development and data usage (Janssen et al., 2020). No prior work has examined generalization, fairness, and explainability together across multiple NLP datasets in one domain. Our study fills this gap with the first comprehensive analysis of these three dimensions across diverse NLP datasets, offering insights that bridge traditionally siloed research directions.

3 Methodology

We use 21 hate speech datasets from MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection (Piot et al., 2024). This selection offers a representative coverage of the hate speech detection landscape through its diversity of platforms (Twitter, Reddit, Gab, Wikipedia), annotation approaches, and targeted domains including gender-based, political, racial, and cyber-bullying content. Table 2 in Appendix A presents a description of each dataset used in the study, along with the source, the original annotation scheme, and the size. Piot et al. (2024) have standardized the heterogeneous annotation schemes by converting all labels into a binary classification of hate speech (positive) versus non-hate speech (negative). We use a Logistic Regression (LR) model with Term Frequency-Inverse Document Frequency (TF-IDF) (Robertson, 2004) and a DistilBert (DB) model (Sanh, 2019), enabling analysis across both inter-

pretable and black-box approaches. LR employs five-fold cross-validation with stratified sampling to maintain consistent class distributions. For DB, we fine-tune the base-uncased weights from HuggingFace (Wolf, 2019) using the AdamW optimizer (Loshchilov et al., 2017) for 3 epochs. In both architectures, we use an 80/20 train-test split. For each dataset, we examine the following: distributional robustness against covariate shift, demographic subgroup performance invariance, and impact on post-hoc explainability. While we expect we could improve predictive performance by experimenting other classifiers, we aim to investigate variations as a function of the training distribution rather than the choice of the classifier. Note, we have selected a minimal yet robust set of analytical tools with high utility across diverse comparative scenarios, as the methodological possibilities for dataset comparison are limitless and could prove counterproductive to navigate. Our framework incorporates well-established metrics in the literature yet remains methodologically flexible, allowing for substitution according to the specific research or operational requirements. We excluded LLMs to maintain our focus on data-centric issues and enable fair comparisons with conventional architectures, as their massive pre-training and transfer learning dynamics would introduce confounding variables. Our findings may still provide valuable insights for LLM fine-tuning dataset selection. Experiments used both personal workstations and a Linux server (40 cores, 125GB RAM).

3.1 Robustness against distribution shift

Machine learning models operate under the closed-world assumptions that the training and inference regimes align. This premise rarely holds in deployment environments, where annotation processes are inherently constrained by incomplete domain expertise, systematic sampling biases, and finite coverage of the target distribution's support (Paullada et al., 2021). Curating datasets often involves multiple degrees of freedom (e.g. source selection, linguistic constraints, perspective samplings, and annotation demographics). Each of them can introduce model degradation: source selection can lead to domain mismatch, linguistic constraints may create artificial patterns that do not generalize, perspective sampling can embed unwanted correlations, and annotation demographics may encode biases in the ground truth. Hence, despite aiming to capture real-world phenomena, datasets become constrained snapshots of the represented field.

The datasets selected in this study aim to represent hate-speech. We aim to measure how well they are designed to do so. For each training distribution, we compute two complementary metrics: (a) the mean cross-domain performance, measured as the average model AUC across all out-domain test sets, and (b) the generalization delta, calculated as the difference between in-distribution test performance and mean cross-domain performance. In doing so, we quantify for each source training distribution, both the absolute cross-domain generalization capacity and the relative performance degradation under distribution shift.

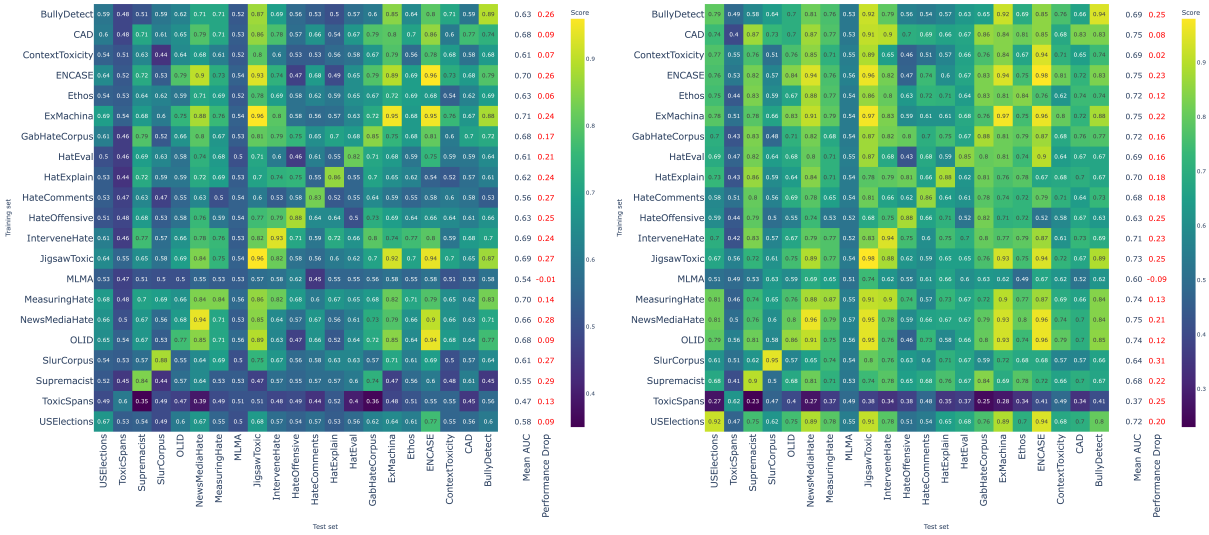
3.2 Classification parity

The decision boundary of a machine learning system is fundamentally shaped by both its positive and negative training observations, where the negative implicitly defines "the rest of the world" (Torralba and Efros, 2011). While datasets must employ compressed representations of this vast instance space, non-representative sampling leads to overconfident classifiers with poor discriminative power. This sampling bias can be particularly problematic when it results in unfair treatment of different demographic groups. We therefore investigate how different training distributions affect model performance across demographic groups. For each source training distribution, we evaluate the resulting trained model using the comprehensive AUC-based metric suite developed by Borkan et al. 2019. The evaluation framework quantifies classification parity through: Subgroup AUC, Background Positive Subgroup Negative (BPSN) AUC, Background Negative Subgroup Positive (BNSP) AUC, Generalized Mean of Bias AUCs (GMB). A detailed description of these metrics can be found in Appendix B. The models are evaluated on the grounds of how much they are able to reduce the unintended bias towards a target community. We conduct our evaluation using the training set of the Jigsaw Unintended Bias in Toxicity Classification competition dataset (Xiao et al.), because it provides explicit identity labels for demographic groups mentioned in each comment. The GMB metric was introduced by the Google Conversation AI Team as part of their Kaggle competition.

3.3 Post-hoc explainability

Recent studies have highlighted that post-hoc explainability methods can be unstable or contradic-

Figure 1: Cross-dataset generalization performance comparison between LR (left) and DB (right) models showing AUC classification scores when training on one dataset (rows) and testing on another (columns), with diagonal cells representing in-domain performance and off-diagonal cells indicating cross-domain generalization capabilities.



tory, either because vulnerable to input perturbations or sensitive to noise or imperceptible artifacts (Ghorbani et al., 2019; Noppel and Wressnegger, 2024; Slack et al., 2020; Dombrowski et al., 2019; Adebayo et al., 2018; Alvarez-Melis and Jaakkola, 2018; Lee et al., 2019). To evaluate and address these stability concerns, researchers need ways to assess the correctness of estimated feature relevances. Assessing the correctness of estimated feature relevances requires a reference "true" influence to compare against. Since this is rarely available, a common approach to measuring the faithfulness of relevance scores with respect to the model they are explaining relies on a proxy notion of importance: observing the effect of removing features on the model's prediction.

We aim to examine how dataset characteristics influence the correctness of post-hoc explainability methods by evaluating feature importance explanations for individual data points using test-time input ablations. The influence of training data on post-hoc explanation faithfulness remains in fact understudied despite its crucial role in model representations, while there is extensive research on model architectures and attribution methods.

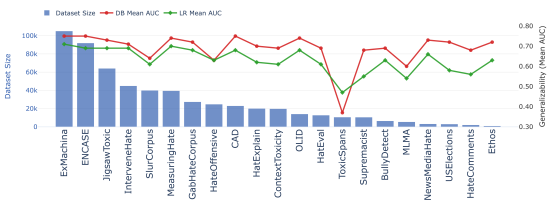
We use the Sufficiency and Comprehensiveness metrics from the ERASER framework (DeYoung et al., 2019) as our evaluation criteria because widely adopted in the literature (Mathew et al., 2020; Carton et al., 2020; Chan et al., 2022; Zhou and Shah, 2022; Wiegreffe and Marasović, 2021). The description of Comprehensiveness and Sufficiency is provided in Appendix C. We keep (for sufficiency) or mask (for comprehensiveness) the top 30% of tokens extracted by the feature importance method as in Sithakoul et al. 2024. We employ these metrics to evaluate explanations generated by SHAP (KernelSHAP, Lundberg, 2017) and LIME (Ribeiro et al., 2016) on both in-distribution samples and out-of-distribution samples from HateXplain (n=500) with SHAP (Mathew et al., 2020).

We hypothesize that increased data complexity, particularly in terms of feature interaction density, leads to reduced faithfulness in LIME explanations due to their local linearity constraints. It impacts SHAP explanations differently through its marginal contribution framework, thus revealing distinct failure modes between the two methods when handling complex linguistic patterns.

4 Results

In this section, we present and analyze the findings from our experiments.

Figure 2: Mean AUC scores by dataset size, comparing LR (green) and DB (red) models.



4.1 Robustness against distribution shift

We evaluate how well a model trained on one dataset generalizes on a representative set of other datasets, compared with its performance on the test set originating from its training distribution. Figures 1 present the cross-dataset generalization performance for the LR (left) and DB (right) models, respectively. The difference in cross-domain performance makes LR a more reliable probe of dataset limitations, as it lacks DB transfer learning advantages. Each row corresponds to training on one dataset and testing on all the others. As expected, both architectures achieve peak performance during in-distribution evaluation. While LR and DB achieve comparable in-domain performance, the LR's learned representations report significantly limited cross-dataset generalization. In LR, *ExMachina* demonstrates the best generalization capability with mean AUC of 0.71, despite its performance drop of 0.24, followed by *MeasuringHate* (mean AUC 0.70, drop 0.14).

To visualize the feature space proximity affecting cross-dataset generalization, we employed UMAP (Becht et al., 2019) on two embedding types: TF-IDF features from LR models and [CLS] token embeddings from fine-tuned DB models. We applied Truncated SVD (Hansen, 1987) to both embedding types (150 components to account for at least 80% explained variance) before projecting to 2D using UMAP (n_neighbors=15, min_dist=0.1) with cosine similarity. We observe correspondence between UMAP semantic representation and cross-dataset generalization metrics. In Figure 3 (left: LR with TF-IDF embeddings), datasets form distinct, island-like clusters, yielding satisfactory in-domain performance but poorer generalization (evident from the stark contrast between diagonal and off-diagonal cells in the heatmap), compared to DB's flowing, interconnected representations that report higher transferability (right). For TF-IDF embeddings, high-performing datasets (*ExMachina*: 0.71 mean AUC, *ENCASE*: 0.70 mean AUC, *JigsawToxic*: 0.69 mean AUC) occupy strategic positions in the embedding space with *JigsawToxic* and *ExMachina* central clusters serving as semantic hubs. *ToxicSpans*' diffuse representation correlates with poor generalization (0.47 mean AUC), while *Supremacist*'s isolation (bright yellow cluster) aligns with limited transferability (0.55 mean AUC). *MLMA*'s scattered distribution across multiple semantic regions might correspond with its cross-domain stability (-0.01 mean AUC performance drop). *NewsMediaHate*'s peripheral positioning explains its significant transferability decline (0.25 drop) despite 0.63 mean AUC. *ENCASE* (0.70 mean AUC) exhibits multiple distinct clusters across the embedding space, suggesting it captures diverse toxic language patterns.

There is a prevailing notion in the literature that increasing the size of the training set might lead to improved model robustness to shift. The LR's marginal improvement with increased training data (Figure 2) suggests that out-of-domain generalization is primarily determined by training-test distributional alignment rather than dataset scale. The performance comparison between LR and DB on individual datasets is reported in Appendix D.

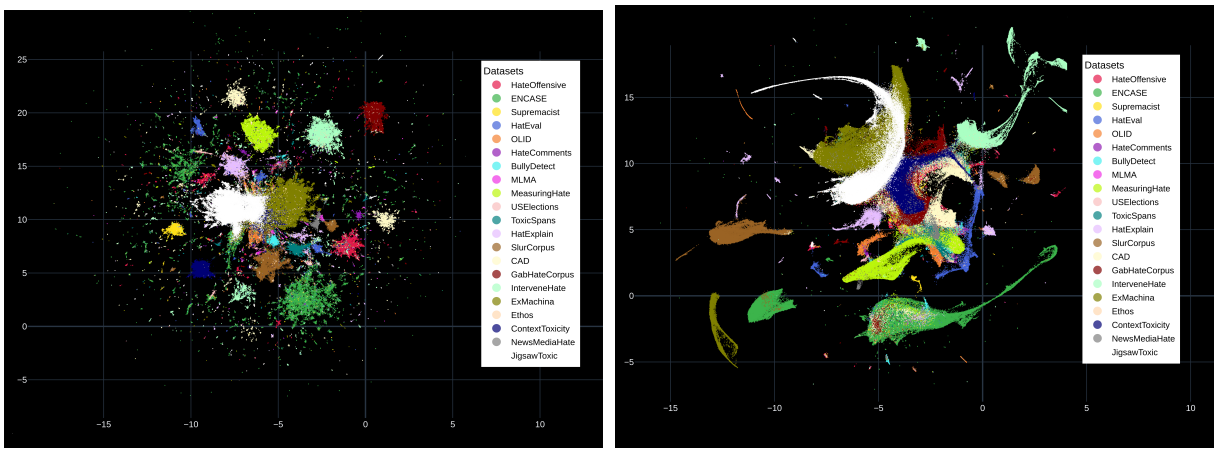
Table 1: Comparison of LR and DB models on the Jigsaw Unintended Bias dataset, reporting AUC and classification parity across demographics (GMB, where higher values indicate more equitable performance across demographics).

Dataset	Logistic Regression		DistilBERT	
	AUC	GMB	AUC	GMB
OLID	0.799	0.560	**0.926**	**0.769**
NewsMediaHate	0.691	0.506	0.870	0.740
ENCASE	0.837	0.624	0.912	0.692
InterveneHate	0.645	**0.654**	0.655	0.507
HateComments	0.593	0.539	0.731	0.621
Ethos	0.582	0.546	0.755	0.630
CAD	0.702	*0.500*	0.786	0.570
BullyDetect	0.674	0.512	0.777	0.573
ExMachina	0.820	0.502	0.877	0.576
GabHateCorpus	0.681	0.516	0.759	0.542
USElections	0.594	*0.500*	0.797	0.545
SlurCorpus	0.554	0.536	0.572	0.545
HatExplain	0.579	0.539	0.658	0.518
HatEval	0.576	*0.500*	0.659	0.519
JigsawToxic	0.770	0.511	0.858	0.530
Supremacist	0.550	0.530	0.695	0.512
MeasuringHate	0.677	0.535	0.750	0.516
MLMA	0.532	*0.500*	0.631	*0.500*
HateOffensive	0.635	0.505	0.639	*0.500*
ContextToxicity	0.671	0.594	0.822	*0.500*
ToxicSpans	*0.458*	*0.500*	*0.391*	*0.500*

4.2 Classification parity

We evaluate model bias across demographic groups using the AUC-based metrics suite from Borkan et al. 2019. Table 1 presents the GMB score of each resulting trained model. The additional fairness metrics (Subgroup AUC, BPSN AUC, and BNSP) AUC are reported in Appendix E. Our analysis reveals consistently low GMB values (0.5-0.7) across all training sets, regardless of their tempo-

Figure 3: UMAP visualization showing cosine similarity relationships between dataset distributions based on (left) TF-IDF embeddings from LR models and (right) token embeddings from fine-tuned DB models.



ral origin, collection methodology, or annotation protocol. This finding has two critical implications. First, traditional classification metrics may obscure significant demographic bias. Models achieving strong predictive performance (AUC $>$ 0.85) simultaneously demonstrate GMB scores approximating random chance ($\approx$ 0.5). Second, this pattern's prevalence across 85.7% of datasets suggests a systematic failure in current dataset construction methods to capture demographic variation in hate speech. Notably, even DB, despite its large-scale pre-training, exhibits similar GMB patterns.

4.3 Post-hoc explainability

Figure 4: Faithfulness metrics of LIME (solid line) and SHAP (dotted line) explanations for in-domain evaluation: comparing sufficiency (top graph, lower values are better) and comprehensiveness (bottom graph, higher values are better) between LR (circle markers) and DB models (triangle markers).

In-domain faithfulness of post-hoc explanations. Figure 4 presents a comparative analysis of SHAP and LIME explanations through sufficiency and comprehensiveness metrics. In line with the literature, we find that linear models tend to achieve better faithfulness metrics compared to transformer-based architectures, with this disparity being particularly pronounced in sufficiency scores. We find that post-hoc explanations do not necessarily have high sufficiency and high comprehensiveness. The most extreme case is DB trained on *ENCASE*, *InterveneHate*, or *SlurCorpus*, which reports good comprehensiveness but poor sufficiency on the same post-hoc explainability method. This discrepancy suggests that the model relies on complex feature interactions rather than independent feature contributions, where removing identified features significantly impacts model confidence but preserving only these features fails to maintain the original prediction.

We observe significant variations in faithfulness across training distributions, independent of the model architecture. Specifically, when controlling for both the architecture and the post-hoc explanation method, the comprehensiveness scores for *InterveneHate* are consistently higher than those for *Supremacist* and *USElections*. We observe variations in faithfulness which persist even in cases where models demonstrate comparable predictive performance across their respective training environments. LR models trained on *JigsawToxic* and *InterveneHate* achieve similar AUC scores (0.95 and 0.93) yet exhibit a more than five-fold difference in comprehensiveness scores (0.12 vs 0.92).

Out-domain faithfulness of post-hoc explanations. We evaluate all models on a common out-of-distribution test set (HateXplain) using SHAP attributions, which demonstrated superior faithfulness to model architectures in our previous analysis. This setup provides a controlled comparison where all models face identical test conditions, allowing us to isolate how different training environments affect explanations faithfulness. Figures 5 and 6 com-

pare the in-domain and out-domain SHAP comprehensiveness and sufficiency scores, respectively, against predictive performance for both the LR and the DB models. To ensure consistent scaling across all models and evaluation settings, both sufficiency and comprehensiveness scores are normalized globally by dividing each value by the maximum absolute value found across all scores, preserving the directionality of each metric (negative for sufficiency with lower being better, positive for comprehensiveness with higher being better). We hypothesize that when a model's predictive performance drops in out-of-domain settings, comprehensiveness and sufficiency scores should correspondingly decrease, as these metrics are based on predictive likelihood which should lower for well-calibrated models (Desai and Durrett, 2020). Out-of-domain evaluation provides a natural setting where model performance degrades, allowing us to test whether faithfulness scores might follow this performance degradation or vary independently when controlling for both the feature attribution method and model architecture.

While AUC scores predictably degrade in out-of-domain settings (green dotted consistently below red), sufficiency scores (Figure 6) improve under domain shift across multiple training datasets, particularly *HateComments* with LR, and *Supremacist*, *GabHateCorpus*, *HateOffensive*, and *JigsawToxic* with DB. This counterintuitive relationship intensifies among DB models (right panel), where out-of-domain sufficiency consistently outperforms its in-domain counterpart. Similarly, comprehensiveness scores (Figure 5) show notably higher values for out-of-domain evaluations in datasets like *Ex-Machina*, *CAD* and *OLID* for both models, despite the degradation in predictive performance.

Statistical analysis reveals distinct patterns in how models trained on different source datasets maintain explanation faithfulness under domain shift. Wilcoxon signed-rank tests show that LR exhibits significant degradation in both sufficiency scores ($\Delta = -0.0220$, $p < 0.001$, $d = 0.31$) and performance ($\Delta = -0.2276$, $p < 0.001$, $d = 0.89$). In contrast, DB maintains consistent sufficiency scores ($\Delta = 0.0000$, $p = 1.000$) despite comparable performance degradation ($\Delta = -0.2062$, $p < 0.001$, $d = 0.84$). Comprehensiveness remains stable across domain shifts for both architectures (LR: $\Delta = -0.0052$, $p = 0.610$; DB: $\Delta = -0.0019$, $p = 0.856$). Notably, we observe no significant correlation between performance drops and metric changes ($\rho = 0.12$, $p = 0.341$), indicating that faithfulness of explanations under domain shift might operate independently from model predictive power. The observed decoupling between performance degradation and explanation faithfulness metrics, might suggest that the underlying learned feature representations might mediate the faithfulness of post-hoc explanations, independent of model performance. Appendix F reports examples of how models trained on different source datasets exhibit marked differences in their SHAP feature attributions when tested on identical out-of-distribution sentences, despite making similarly high-confidence hate speech predictions.

5 Discussion

We analyzed how learned representations in hate speech detection models are shaped by 21 different training datasets, examining robustness to distribution shifts, demographic representation, and post-hoc explainability. Our findings aim to help practitioners assess dataset suitability for their specific applications and understand potential downstream limitations of their model.

> **Observation 1:** *Training distributions exhibit inherent divergence from one another, as evidenced by consistent performance degradation in cross-domain evaluation, despite shared semantics and annotation frameworks.*

Machine learning models operate under the assumption of distributional alignment between training and test distributions - an assumption our cross-domain experiments systematically invalidate. We demonstrate substantial distributional heterogeneity, manifesting in significant performance degradation when models are evaluated on distributions different from their training data. This heterogeneity persists even among datasets sharing the same domain objectives and annotation frameworks, highlighting fundamental limitations in dataset curation. This distributional heterogeneity is confirmed in the UMAP visualization, where datasets embeddings often form distinct, isolated clusters in the semantic space despite addressing similar objectives.

> **Observation 2:** *The simultaneous optimisation of distribution robustness and demographic fairness remains elusive.*

Our empirical evaluation demonstrates that 85.7% of datasets exhibit GMB performance at random chance (0.5), with models failing to simul-

Figure 5: Comparison of in-domain and out-domain SHAP comprehensiveness scores against AUC scores for DB and LR.

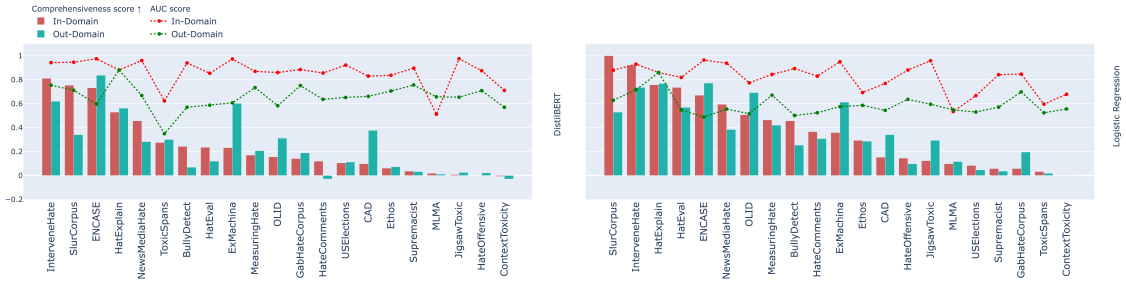
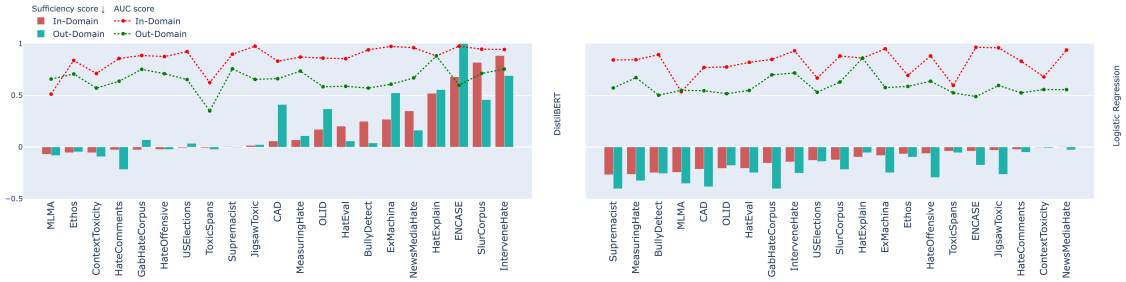


Figure 6: Comparison of in-domain and out-domain SHAP sufficiency scores against AUC scores for DB and LR.



taneously achieve predictive accuracy and demographic fairness. This pattern manifests in two distinct outcomes: models either maintain predictive accuracy while violating fairness criteria, or fail at both metrics. While this could suggest representation gaps in training data (covariate shift), the observed performance patterns might equally stem from systematic label bias (concept drift) in cross-cultural interpretation.

> **Observation 3:** *Post-hoc explanation faithfulness demonstrates complex, non-trivial dependencies on learned representations, model architectures and attribution methods, while remarkably maintaining or improving despite significant performance degradation in out-of-domain settings.*

Post-hoc explainability methods, when evaluated on models trained and tested on the same distribution (in-domain), exhibit volatility independent of feature attribution methods and model architectures. This instability manifests even across models with comparable predictive performance. In cross-distribution evaluation (out-domain), where multiple models trained on different datasets are tested against a common distribution, we observe that while predictive performance degrades predictably, explanation faithfulness metrics show inconsistent and often contradictory responses. The absence of correlation between faithfulness metric changes and performance degradation suggests that the learned feature representations might mediate the faithfulness of post-hoc explanations, independent of the model predictive power. This crucial disconnect challenges the methods reliability in practical applications presenting distribution shifts.

6 Conclusion

Rather than advocating for larger and enhanced datasets - an approach that reinforces the field's fixation on scale – we aimed to foster a deeper reflection on the impact of dataset selection under the lens of model behavior. While achieving high AUC on individual hate speech benchmarks might suggest progress, our analysis of learned representations across 21 datasets reveals: pervasive distributional divergence evidenced by cross-domain performance degradation, the inability to simultaneously ensure robustness and demographic fairness, and complex dependencies with post-hoc explainability faithfulness.

7 Limitations

While numerous metrics exist for evaluating model behavior, we deliberately restricted our focus to a core set that are both widely validated in literature and directly relevant to our research objectives. The sufficiency and comprehensiveness metrics employ a fixed threshold for feature masking, which may not be optimal across all cases and warrants exploration of additional thresholds. These metrics

also require producing counterfactual inputs that are inherently out-of-distribution to models. Our concerns about this methodological constraint echo those raised in prior work (Hase et al., 2021). We maintained methodological consistency across all comparisons, ensuring that even if our chosen metrics have inherent limitations, these limitations affect all input distributions equally. This means any ablation-based artifacts present when tested against *HateXplain* would impact all distributions in the same way. Finally, our model selection was limited to traditional classifiers and pre-trained transformers like DB, deliberately excluding LLMs, as their billion-scale parameter spaces and large-scale pre-training would have confounded our primary objective of isolating dataset-specific effects on model behavior. As outside the scope of this paper, in future work we will investigate the mechanisms behind high levels of post-hoc explainability faithfulness observed in conditions of model degradation on out-of-distribution sentences. Preliminary experiments, reported in Appendix G, seem to suggest that when models encounter out-of-distribution inputs, they resort to simpler heuristics—using fewer features with more concentrated importance. This leads to increased explanation faithfulness despite degraded model performance, as explanation methods more accurately capture these simplified decision patterns rather than more complex reasoning exhibited in-distribution. As future work, we plan to conduct controlled experiments that will help establish causal mechanisms underlying the correlations identified in our current study.

7.1 Ethical Considerations

This study examines hate speech dataset variations through three model desiderata, recognizing that performance differences often reflect legitimate contextual distinctions rather than methodological flaws. Examples appear without identifying metadata, and research received institutional ethics approval.

8 Acknowledgments

We used AI language models for proofreading portions of the paper.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. In Advances in Neural Information Processing Systems, pages 9505–9515.

Saad Almohaimeed, Saleh Almohaimeed, Ashfaq Ali Shafin, Bogdan Carbunar, and Ladislau Bölöni. 2023. THOS: A benchmark dataset for targeted hate and offensive speech. In Workshop on Data-centric machine learning.

David Alvarez-Melis and Tommi S Jaakkola. 2018. On the robustness of interpretability methods. In ICML Workshop on Human Interpretability in Machine Learning.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In Proceedings of the 42nd international acm sigir conference on research and development in information retrieval, pages 45–54.

Adrian Arnaiz-Rodriguez and Nuria Oliver. 2024. Towards algorithmic fairness by means of instance-level data re-weighting based on Shapley values. In Workshop on Data-centric machine learning.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In International conference on machine learning, pages 233–242. PMLR.

Varun Babbar, Zhicheng Guo, and Cynthia Rudin. 2024. What is different between these datasets? arXiv preprint arXiv:2403.05652.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In Proceedings of the 26th international conference on World Wide Web companion, pages 759–760.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 54–63. ACL.

Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using umap. Nature biotechnology, 37(1):38–44.

Tazeek Bin Abdur Rakib and Lay-Ki Soon. 2018. Using the reddit corpus for cyberbully detection. pages 180–189. Springer.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In Companion proceedings of the 2019 world wide web conference, pages 491–500.

Samuel Carton, Anirudh Rathore, and Chenhao Tan. 2020. Evaluating and characterizing human rationales. arXiv preprint arXiv:2010.04736.

Chun Sik Chan, Huanqi Kong, and Guanqing Liang. 2022. A comparative study of faithfulness metrics for model interpretability methods. arXiv preprint arXiv:2204.05514.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. arXiv preprint arXiv:1905.12516.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. Proceedings of the ICWSM 2017, 11(1):512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In Proceedings of the ALW2 2018. ACL.

Junwei Deng and Jiaqi Ma. 2024. Computational copyright: Towards a royalty model for AI music generation platforms. In Workshop on Data-centric machine learning.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. arXiv preprint arXiv:2003.07892.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. arXiv preprint arXiv:1911.03429.

DMLR. 2024. Call for papers DMLR 2024. https://dmlr.ai/cfp-icml24/. Accessed: 2025-02-15.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. arXiv preprint arXiv:1906.07983.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pages 214–226. ACM.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of nlp leaderboards. arXiv preprint arXiv:2009.13888.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 259–268. ACM.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, others, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. Proceedings of the ICWSM 2018, 12(1).

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 3681–3688.

Lukas Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection. In Proceedings of the WASSA 2021, pages 171–180. ACL.

Sarthak Gupta, Pranav Priyadarshi, and Manish Gupta. 2023. Hateful comment detection and hate target type prediction for video comments. In Proceedings of the CIKM 2023, CIKM '23, pages 3923–3927. ACM.

Per Christian Hansen. 1987. The truncated svd as a method for regularization. BIT Numerical Mathematics, 27:534–553.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems, pages 3315–3323.

Peter Hase, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. Advances in neural information processing systems, 34:3650–3666.

Marijn Janssen, Paul Brous, Elsa Estevez, Luis S Barbosa, and Tomasz Janowski. 2020. Data governance: Organizing data for trustworthy artificial intelligence. Government information quarterly, 37(3):101493.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2019. Fantastic generalization measures and where to find them. arXiv preprint arXiv:1912.02178.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, others, and Morteza Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. Language Resources and Evaluation, 56(1):79–108.

Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Caterina von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application.

Ravin Kohli, Matthias Feurer, Katharina Eggensperger, Bernd Bischl, and Frank Hutter. 2024. Towards quantifying the effect of datasets for benchmarking: A look at tabular machine learning. In Workshop on Data-centric machine learning.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In Proceedings of the WOAH 2020, pages 138–149, Online. ACL.

Eunjin Lee, David Braines, Mitchell Stiffler, Adam Hudler, and Daniel Harborne. 2019. Developing the sensitivity of lime for better machine learning explanation. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, volume 11006, page 1100610. International Society for Optics and Photonics.

Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. arXiv preprint arXiv:1711.05101, 5.

Scott Lundberg. 2017. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI 2020.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, pages 220–229.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. Complex & Intelligent Systems, 8(6):4663–4678.

Rafael Mosquera Gómez, Julian Eusse, Juan Ciro, Daniel Galvez, Ryan Hileman, Kurt Bollacker, and David Kanter. 2023. Speech Wikimedia: A 77 language multilingual speech dataset. In Workshop on Data-centric machine learning.

Maximilian Noppel and Christian Wressnegger. 2024. Sok: Explainable machine learning in adversarial environments. In 2024 IEEE Symposium on Security and Privacy (SP), pages 2441–2459. IEEE.

OECD. 2024. Recommendation of the council on artificial intelligence. Accessed: 16 May 2025.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In Proceedings of the EMNLP-IJCNLP 2019, pages 4675–4684. ACL.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11).

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter?

John Pavlopoulos, Jeffrey Sorensen, Léa Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In Proceedings of the SemEval 2021, pages 59–69. ACL.

Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. L3CubeMahaSent-MD: A multi-domain Marathi sentiment analysis dataset and transformer models. In Workshop on Data-centric machine learning.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection. In Proceedings of the International AAAI Conference on Web and Social Media, volume 18, pages 2025–2039.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In Proceedings of the EMNLP-IJCNLP 2019, pages 4755–4764. ACL.

José Ribeiro, Raíssa Silva, Lucas Cardoso, and Ronnie Alves. 2021. Does dataset complexity matters for model explainers? In 2021 IEEE International Conference on Big Data (Big Data), pages 5257–5265. IEEE.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pages 1135–1144.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation, 60(5):503–520.

Andrea Romei and Salvatore Ruggieri. 2014. A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review, 29(05):582–638.

Pranav Sachdeva, Ricardo Barreto, Geoff Bacon, Alexander Sahn, Caterina von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In Proceedings of the LREC 2022, pages 83–94. ELRA.

Joni Salminen, Hind Almerekhi, Milos Milenkovic, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard Jansen. 2018. Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. Proceedings of the ICWSM 2018, 12(1).

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pages 1–15.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Rajat Shinde, Sujit Roy, Christopher E Phillips, Aman Gupta, Aditi Sheshadri, Manil Maskey, and Rahul Ramachandran. 2024. WINDSET: Weather insights and novel data for systematic evaluation and testing. In Workshop on Data-centric machine learning.

Samuel Sithakoul, Sara Meftah, and Clément Feutry. 2024. Beexai: Benchmark to evaluate explainable ai. In World Conference on Explainable Artificial Intelligence, pages 445–468. Springer.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In Conference on Artificial Intelligence, Ethics, and Society (AIES).

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In CVPR 2011, pages 1521–1528. IEEE.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset. In Proceedings of the NAACL 2021, pages 2289–2303. ACL.

Artem Vysogorets and Julia Kempe. 2024. Towards robust data pruning. In Workshop on Data-centric machine learning.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable natural language processing. arXiv preprint arXiv:2102.12060.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale.

Yao Xiao, Yaoyao Chang, Cheng Peng, Siyu Li, and Zhiyu Yuan. Jigsaw unintended bias in toxicity classification.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. 2024. Generalized out-of-distribution detection: A survey. International Journal of Computer Vision, pages 1–28.

Rubing Yang, Jialin Mao, and Pratik Chaudhari. 2022. Does the data induce capacity control in deep learning? In International Conference on Machine Learning, pages 25166–25197. PMLR.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In Proceedings of the NAACL 2019, pages 1415–1420. ACL.

Dorothy Zhao, Alice Xiang, Jerone T A Andrews, and Orestis Papakyriakopoulos. 2024. Measuring diversity in datasets. In Workshop on Data-centric machine learning.

Yilun Zhou and Julie Shah. 2022. The solvability of interpretability evaluation metrics. arXiv preprint arXiv:2205.08696.

Indre Zliobaite. 2015. A survey on measuring indirect discrimination in machine learning. arXiv preprint arXiv:1511.00148.

9 NLP Checklist

A1. Did you describe the limitations of your work? Yes. Please refer to Section 7.

A2. Did you discuss any potential risks of your work? Yes. We have discussed some ethical considerations in Section 7.1.

B. Did you use or create scientific artifacts? Yes, we used existing scientific artifacts (datasets, pre-trained models, evaluation metrics).

B1. Did you cite the creators of artifacts you used? Yes. Please refer to Section 3.

B2. Did you discuss the license or terms for use and/or distribution of any artifacts? Yes. Please refer to Section 3.

B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)? Our use of the MetaHate dataset follows its intended research purpose, accessed through a signed terms of use agreement. Any derivatives from our work maintain the original research-only restrictions and cannot be used outside research contexts.

B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it? The datasets

used contain offensive language. The sources are publicly available, however, to avoid any distressing feeling to our readers we avoided presenting and cite content in the full body of the paper that can affect the readers.

B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.? The datasets are explained in detail by the authors of the MetaHate paper. The descriptions in this paper include only what is necessary to this work.

B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created? Yes, please refer to Sections 3 and 4.

C. Did you run computational experiments? Yes.

C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used? Yes. Please refer to Section 3.

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values? Yes. We did not perform hyperparameter tuning.

C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run? Yes. We report performance results in Section 4 and Appendix D.

C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, Spacy, ROUGE, etc.), did you report the implementation, model, and parameter settings used? Yes. Please refer to Section 3.

D. Did you use human annotators (e.g., crowdworkers) or research with human participants? No.

D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.? N/A.

D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)? N/A.

D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A.

D4. Was the data collection protocol approved (or determined exempt) by an ethics review board? N/A.

D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data? N/A.

E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing? Yes. We used AI language models for proofreading portions of the manuscript to check for grammatical errors and clarity.

E1. Did you include information about your use of AI assistants? Yes. Please refer to Section 8.

A Datasets used for experimental evaluation

Dataset	Size	Description	Original Annotation	Source	References
Binary Classification					
Hateval 2019	12,747	Hate speech against women and immigrants	Hate, Non-hate	Twitter	Basile et al., 2019
OLID 2019	14,052	Hierarchical offensive language	Hate, Non-hate	Twitter	Zampieri et al., 2019
US 2020 Elections	2,999	Political hate speech	Hate, Non-hate	Twitter	Grimminger and Klinger, 2021
BullyDetect 2018	6,562	Cyberbullying	Cyberbullying, No cyberbullying	Reddit	Bin Abdur Rakib and Soon, 2018
Intervene Hate 2019	45,170	Counter-speech and hate speech	Hate, Non-hate	Reddit, Gab	Qian et al., 2019
Hate in Online News	3,214	News comments	Hate, Non-hate	Facebook	Salminen et al., 2018
Supremacist 2018	10,534	White supremacist content	Hate, Non-hate	Stormfront	de Gibert et al., 2018
Gab Hate Corpus	27,434	Hate speech	Assault on Human Dignity / No	Gab	Kennedy et al., 2022
HateComments 2023	2,070	Hate speech	Hate, Non-hate	YouTube	Gupta et al., 2023
Ex Machina 2016	115,705	Toxicity detection	Attack, No Attack	Wikipedia	Wulczyn et al., 2016
Context Toxicity 2020	19,842	Context-aware toxicity	Toxic, No Toxic	Wikipedia	Pavlopoulos et al., 2020
Multi-class / Multi-label Classification					
Hate Offensive 2017	24,783	Offensive language	Hate Speech, Offensive, Neither	Twitter	Davidson et al., 2017
ENCASE 2018	91,950	Cyberbullying and hate speech	Abusive, Normal, Spam, Hateful	Twitter	Founta et al., 2018
MLMA 2019	5,593	Multilingual hate speech	Multiple abuse categories	Twitter	Ousidhoum et al., 2019
HateXplain 2020	20,109	Explainable hate speech	Hate, Offensive, Normal	Twitter, Gab	Mathew et al., 2020
Slur Corpus 2020	39,960	Slur-based hate speech	Multiple slur categories	Reddit	Kurrek et al., 2020
CAD 2021	23,060	Contextual abuse	Multiple abuse types	Reddit	Vidgen et al., 2021
Severity Scale					
Measuring Hate 2020-22	39,565	Linear hate speech scale	Severity scale	Twitter, Reddit, YouTube	Kennedy et al., 2020; Sachdeva et al., 2022
ETHOS 2020	998	Multi-target hate speech	Severity scale	Reddit, YouTube	Mollas et al., 2022
Span-level Annotation					
Toxic Spans 2021	10,621	Token-level toxicity	Span-level annotation	Comments	Pavlopoulos et al., 2021

Table 2: Description of the dataset adopted for the experimental evaluation.

Note: Datasets are grouped by classification type. For a comprehensive description of each dataset, please refer to Piot et al., 2024. While the original Toxic Spans 2021 dataset (Pavlopoulos et al., 2021) identified specific text segments indicating toxicity, in MetaHate (Piot et al., 2024) the authors have standardized its format to match other datasets, providing binary classifications of whether comments contain hate speech or not. For MLMA 2019, they (Piot et al., 2024) have selected only text in English.

B Classification parity metrics

- Subgroup AUC: We restrict the data set to only the examples that mention the specific identity subgroup. A low value in this metric means the model does a poor job of distinguishing between toxic and non-toxic comments that mention the identity.

- BPSN AUC: We restrict the test set to the non-toxic examples that mention the identity and the toxic examples that do not. A low value in this metric means that the model confuses non-toxic examples that mention the identity with toxic examples that do not, likely meaning that the model predicts

higher toxicity scores than it should for non-toxic examples mentioning the identity.

- BNSP AUC: We restrict the test set to the toxic examples that mention the identity and the non-toxic examples that do not. A low value here means that the model confuses toxic examples that mention the identity with non-toxic examples that do not, likely meaning that the model predicts lower toxicity scores than it should for toxic examples mentioning the identity.

- GMB AUC: This metric was introduced by the Google Conversation AI Team as part of their Kaggle competition. [1] This metric combines the per-identity Bias AUCs into one overall measure as:

$$M_p(m_s) = \left(\frac{1}{N} \sum_{s=1}^{N} m_s^p \right)^{\frac{1}{p}} \tag{1}$$

where:

  - M_p = the p^{th} power-mean function
  - m_s = the bias metric m calculated for subgroup s
  - N = number of identity subgroups (10)

We use $p = -5$ as was done in the competition.

C Faithfulness metrics

- Comprehensiveness represents the impact of replacing most important rationales by a baseline. For each input x_i, we construct a contrast example $\tilde{x}_i = x_i \setminus r_i$ by removing the predicted rationales r_i. Let $m(x_i)_j$ denote the prediction probability assigned by model m to class j for the original input. Comprehensiveness is defined as *comprehensiveness* $= m(x_i)_j - m(x_i \setminus r_i)_j$, where $m(x_i \setminus r_i)_j$ is the prediction probability for the contrast example. When important rationales are removed, we expect the model's confidence to decrease, yielding a higher comprehensiveness score that indicates more faithful interpretations (DeYoung et al., 2019).

- Sufficiency represents the impact of adding most important features to a baseline in the predictive behavior. We measure sufficiency as *sufficiency* $= m(x_i)_j - m(r_i)_j$, where $m(r_i)_j$ is the prediction probability when only rationales are provided. A lower sufficiency score implies the rationales contain essential information for the model's prediction, suggesting more faithful interpretations (DeYoung et al., 2019).

[1] Jigsaw Unintended Bias in Toxicity Classification competition: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

D Performance comparison of LR and DB across different hate speech datasets

Dataset	F1		AUROC		Precision		Recall		Accuracy		Bal. Acc	
	LR	DB	LR	DB	LR	DB	LR	DB	LR	DB	LR	DB
MLMA	0.433	0.433	0.534	0.512	0.000	0.000	0.000	0.000	0.763	0.765	0.499	0.500
HatEval	0.724	0.760	0.819	0.854	0.719	0.734	0.610	0.696	0.739	0.769	0.720	0.759
NewsMediaHate	**0.848**	**0.886**	0.938	**0.960**	**0.937**	**0.941**	0.878	**0.928**	0.871	**0.907**	0.865	**0.890**
MeasuringHate	0.698	0.744	0.844	0.870	0.718	0.716	0.407	0.523	0.802	0.820	0.675	0.725
HateOffensive	0.566	0.537	0.881	0.875	0.684	0.640	0.091	0.056	**0.945**	**0.944**	0.544	0.527
ToxicSpans	0.479	0.479	0.596	0.623	0.918	0.918	1.000	1.000	0.918	0.918	0.500	0.500
CAD	0.574	0.691	0.769	0.830	0.791	0.758	0.144	0.336	0.831	0.854	0.568	0.656
HateComments	0.725	0.782	0.830	0.856	0.688	**0.805**	0.759	0.718	0.725	0.785	0.727	0.781
Supremacist	0.535	0.637	0.842	0.897	0.727	0.762	0.069	0.207	0.895	0.906	0.533	0.599
SlurCorpus	0.805	**0.882**	0.880	0.946	0.806	0.894	0.816	0.874	0.805	0.882	0.805	0.882
HatExplain	0.792	0.795	0.860	0.881	0.772	0.859	0.739	0.653	0.798	0.809	0.790	0.787
ExMachina	0.826	0.868	0.950	0.973	0.880	0.920	0.568	0.657	0.932	0.947	0.778	0.824
ContextToxicity	0.497	0.497	0.678	0.711	0.000	0.000	0.000	0.000	0.988	0.988	0.500	0.500
ENCASE	0.920	0.927	**0.964**	**0.975**	0.891	0.886	0.875	0.905	0.937	0.942	**0.918**	**0.930**
Ethos	0.624	0.736	0.693	0.837	0.500	0.620	0.515	0.721	0.660	0.755	0.625	0.747
USElections	0.469	0.762	0.667	0.922	0.000	0.811	0.000	0.435	0.885	0.923	0.500	0.711
JigsawToxic	0.767	0.835	0.959	0.975	0.857	0.743	0.408	0.641	0.962	0.967	0.702	0.814
OLID	0.672	0.764	0.774	0.860	0.775	0.756	0.378	0.604	0.753	0.800	0.661	0.752
BullyDetect	0.758	0.839	0.893	0.940	0.855	0.736	0.489	0.815	0.834	0.867	0.728	0.851
GabHateCorpus	0.559	0.666	0.847	0.885	0.737	0.688	0.090	0.255	0.920	0.927	0.543	0.622
InterveneHate	0.888	0.904	0.930	0.943	0.909	0.899	0.825	0.874	0.893	0.907	0.883	0.902

Table 3: Performance comparison between Logistic Regression (LR) and DistilBERT (DB) models across 21 hate speech datasets. Metrics include F1 score, Area Under the Receiver Operating Characteristic curve (AUROC), Precision, Recall, Accuracy, and Balanced Accuracy. Bold values indicate the highest performance across datasets for each model type.

E Subgroup, BPSN, and BNSP AUC metrics of LR and DB across different hate speech datasets

Dataset	Subgroup AUC		BPSN AUC		BNSP AUC	
	LR	DB	LR	DB	LR	DB
MLMA	0.500	0.500	0.499	0.500	0.501	0.500
HatEval	0.544	0.524	0.498	0.503	0.561	0.532
NewsMediaHate	0.590	**0.698**	**0.498**	**0.743**	0.646	**0.732**
MeasuringHate	0.504	0.509	0.510	0.523	0.504	0.508
HateOffensive	0.503	0.500	0.499	0.500	0.505	0.500
ToxicSpans	0.500	0.500	0.500	0.500	0.500	0.500
CAD	0.550	0.571	**0.486**	0.547	0.572	0.591
HateComments	0.537	0.612	**0.483**	0.600	0.583	0.636
Supremacist	0.507	0.515	**0.481**	0.506	0.528	0.517
SlurCorpus	0.553	0.558	0.574	0.641	**0.477**	**0.432**
HatExplain	0.549	0.527	0.505	0.501	0.544	0.531
ExMachina	0.541	0.547	0.579	0.606	0.542	0.547
ContextToxicity	0.500	0.500	0.500	0.500	0.500	0.500
ENCASE	**0.629**	**0.667**	0.573	**0.696**	**0.666**	**0.686**
Ethos	0.560	**0.646**	0.500	0.547	0.589	**0.705**
USElections	0.500	0.534	0.500	0.556	0.500	0.535
JigsawToxic	0.502	0.514	0.510	0.547	0.502	0.514
OLID	**0.633**	**0.710**	0.609	**0.769**	**0.693**	**0.764**
BullyDetect	0.506	0.546	0.521	0.600	0.504	0.546
GabHateCorpus	0.528	0.577	**0.484**	**0.455**	0.544	0.613
InterveneHate	0.517	0.509	0.495	0.504	0.526	0.510

Table 4: Comparison of fairness metrics (Subgroup, BPSN, and BNSP AUC) for Logistic Regression (LR) and Debiased (DB) models across 21 hate speech datasets. Values in **bold** highlight notable performance (>0.65 or <0.49).

F Impact of source training data on features attribution

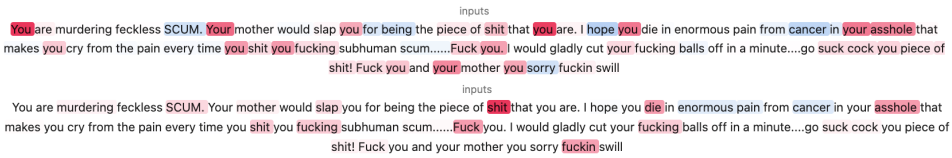


Figure 7: SHAP explanation of the LR model trained on *BullyDetect* (above) and *ENCASE* (below) and tested on the same out of distribution sentence. The scores relate to the predicted probability of the positive class (hate), namely, PP = 1 (above) and PP = 0.99 (below)

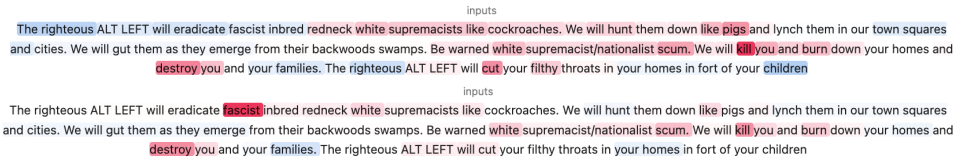


Figure 8: SHAP explanation of the LR model trained on *OLID* (above) and *Ethos* (below) and tested on the same out of distribution sentence. The scores relate to the predicted probability of the positive class (hate), namely, PP = 0.96 (above) and PP = 0.94 (below)

G Post-hoc explainers and out-of-distribution-induced heuristic simplification

We conduct a preliminary experiment to investigate how distribution shift affects post-hoc explainability faithfulness. We used LR trained on *ExMachina*, and compared SHAP explanations between in-distribution (*ExMachina*) and out-of-distribution (*HateXplain*) inputs (n=50 each). We chose *ExMachina* specifically because it demonstrates higher faithfulness scores (both comprehensiveness and sufficiency) when evaluating out-of-domain inputs compared to in-domain inputs. Results revealed that explanations on out-of-distribution inputs relied on significantly fewer significant features (464 vs. 1,210), exhibited higher importance concentration (Gini coefficient 0.994 vs. 0.985), and lower feature entropy (6.83 vs. 8.06). We also observe a dramatic reduction in feature diversity, with out-of-distribution inputs using only 58.28 average unique features per sample compared to 205.92 for in-distribution data. Moreover, there is a a minimal semantic overlap between distributions—only 20% of top features were shared between in-distribution and out-of-distribution inputs. These findings might indicate that when encountering unfamiliar inputs, models resort to simplified decision heuristics that explanation methods can more accurately capture, creating misleading impressions of explanation reliability.